International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: http://www.tandfonline.com/loi/hihc20

The System Usability Scale: Past, Present, and Future

James R. Lewis

To cite this article: James R. Lewis (2018): The System Usability Scale: Past, Present, and Future, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2018.1455307

To link to this article: https://doi.org/10.1080/10447318.2018.1455307

Published online: 30 Mar 2018.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2018.1455307

Taylor & Francis
Taylor & Francis Group

The System Usability Scale: Past, Present, and Future

James R. Lewis

IBM Corporation, Armonk, New York, USA

ABSTRACT

The System Usability Scale (SUS) is the most widely used standardized questionnaire for the assessment of perceived usability. This review of the SUS covers its early history from inception in the 1980s through recent research and its future prospects. From relatively inauspicious beginnings, when its originator described it as a "quick and dirty usability scale," it has proven to be quick but not "dirty." It is likely that the SUS will continue to be a popular measurement of perceived usability for the foreseeable future. When researchers and practitioners need a measure of perceived usability, they should strongly consider using the SUS.

KEYWORDS

Perceived usability; standardized usability scale; System Usability; SUS

1. Introduction

1.1 What is the System Usability Scale?

The System Usability Scale (SUS) is a widely used standardized questionnaire for the assessment of perceived usability. Sauro and Lewis (2009) reported that the SUS accounted for 43% of post-study questionnaire usage in industrial usability studies. Google Scholar citations (examined 3/13/2018) showed 5,664 citations for the paper that introduced the SUS (Brooke, 1996). In its standard (most often used) form, the SUS has 10 five-point items with alternating positive and negative tone (see Figure 1).

1.2. Where did the SUS come from?

The early 1980s saw a dramatic increase in the application of human factors psychology to the design and evaluation of office and personal computer systems. It became clear that for commercial computer products, a focus only on objective usability (effectiveness and efficiency) was insufficient – it was also important to assess perceived usability (ISO, 1998). There were several existing standardized questionnaires for the assessment of user satisfaction with systems (LaLomia & Sidowski, 1990), but they were not designed for the assessment of usability following participation in task-based usability tests.

Several researchers attempted to fill that void with the publication of standardized usability questionnaires in the late 1980s, some developed at universities and others at major corporations. Many of these are still in use, including:

- The Questionnaire for User Interaction Satisfaction (QUIS) (Chin, Diehl, & Norman, 1988) – University of Maryland, College Park
- The Software Usability Measurement Inventory (SUMI) (Kirakowski & Corbett, 1993; McSweeney, 1992) – University College, Cork
- The Post-Study System Usability Questionnaire (PSSUQ) and its non-lab variant, the Computer Systems Usability Questionnaire (CSUQ) (Lewis, 1990, 1992, 1995, 2002) – International Business Machines Corporation
- The SUS (Brooke, 1996) – Digital Equipment Corporation

The SUS was the last of these to be published, but it might have been the first to have been developed. In a retrospective of the SUS, its originator, John Brooke, wrote (2013, p. 29):

> [In 1984] as part of a usability engineering program, I developed a questionnaire – the System Usability Scale (SUS) – that could be used to take a quick measurement of how people perceived the usability of computer systems on which they were working. This proved to be an extremely simple and reliable tool for use when doing usability evaluations, and I decided, with the blessing of engineering management at Digital Equipment Co. Ltd (DEC; where I developed SUS), that it was probably something that could be used by other organizations (the benefit for us being that if they did use it, we potentially had something we could use to compare their systems against ours). So, in 1986, I made SUS freely available to a number of colleagues, with permission to pass it on to anybody else who might find it useful, and over the next few years occasionally heard of evaluations of systems where researchers and usability engineers had used it with some success. Eventually, about a decade after I first created it, I contributed a chapter describing SUS to a book on usability engineering in industry (Brooke, 1996). Since then … it has been incorporated into commercial usability evaluation toolkits such as Morae, and I have recently seen several publications refer to it as an "industry standard" – although it has never been through any formal standardization process.

There are no fees required to use the SUS. "The only prerequisite for its use is that any published report should acknowledge the source of the measure" (Brooke, 1996, p. 194). Thus, researchers who use the SUS should acknowledge Brooke (1996) as the source in internal reports and external publications.

CONTACT James R. Lewis jimlewis@us.ibm.com

	The System Usability Scale Standard Version		Strongly Disagree 1	2	3	4	Strongly Agree 5
1	I think that I would like to use this system frequently.		o	o	o	o	o
2	I found the system unnecessarily complex.		o	o	o	o	o
3	I thought the system was easy to use.		o	o	o	o	o
4	I think that I would need the support of a technical person to be able to use this system.		o	o	o	o	o
5	I found the various functions in this system were well integrated.		o	o	o	o	o
6	I thought there was too much inconsistency in this system.		o	o	o	o	o
7	I would imagine that most people would learn to use this system very quickly.		o	o	o	o	o
8	I found the system very awkward to use.		o	o	o	o	o
9	I felt very confident using the system.		o	o	o	o	o
10	I needed to learn a lot of things before I could get going with this system.		o	o	o	o	o

Figure 1. The standard SUS.

Item 8 shown with "awkward" in place of the original "cumbersome" (Bangor et al., 2008; Finstad, 2006)

1.3. Scoring the SUS

The standard approach to scoring the SUS is somewhat complicated due to the alternating tone of the items and to an early decision to manipulate the score to range from 0 to 100. Conceptually, the first scoring step is to convert raw item scores to adjusted scores (also known as "score contributions") that range from 0 (poorest rating) to 4 (best rating), with that adjustment differing for the odd- and even-numbered items (respectively, the positive- and negative-tone items). The scoring system of the SUS requires ratings for all 10 items, so if a respondent leaves an item blank, it should be given a raw score of 3 (the center of the five-point scale). For the odd-numbered items, subtract 1 from the raw score, and for the even-numbered items, subtract the raw score from 5. Compute the sum of the adjusted scores, then multiply by 2.5 to get the standard SUS score. The following equation shows a more concise way to compute a standard SUS score from a set of raw item ratings:

[1] SUS = 2.5(20 + SUM(SUS01,SUS03,SUS05,SUS07,SUS09) − SUM(SUS02,SUS04,SUS06,SUS08,SUS10))

1.4. Assessing the quality of standardized questionnaires

Before reviewing the research on the SUS, this section offers a summary of the psychometric methods used to assess the quality of standardized questionnaires, focusing on validity (the extent to which a questionnaire measures what it claims to measure), reliability (its consistency of measurement), and sensitivity (the extent to which independent variables affect measurement). The focus in this section is on the methods of classical test theory (CTT, Nunnally, 1978) rather than item response theory (Embretson & Reise, 2000) because CTT has been the prevailing methodology in the development of standardized usability questionnaires.

Content validity

The first consideration in assessing a questionnaire is whether it has valid content, in other words, whether its items are relevant and representative of what it is intended to measure. There are no statistical tests for this type of validity because it depends on rational rather than empirical assessment of the source of the items. Creation of items by domain experts or selection from a literature review of existing questionnaires in the target or related domain are typically taken as evidence supporting the claim of content validity.

Construct validity

Construct validity is assessed by examining the extent to which questionnaire items align with the underlying constructs of interest. Because Likert (summated) scales are more reliable than single-item scores and it is easier to interpret and present a smaller number of scores, the most commonly used analytical method is to conduct a factor analysis to determine if there is a statistical basis for the formation of measurement scales based on factors. Generally, a factor analysis requires a minimum of five participants per item to ensure stable factor estimates (Nunnally, 1978). There are several methods for estimating the number of factors in a set of scores when conducting exploratory analyses, including discontinuity and parallel analysis (Cliff, 1987; Coovert & McNelis, 1988; O'Connor, 2000). This step in the questionnaire development process includes examination of which items most strongly load on the factor(s) of interest so poorer-performing items can be removed. Once previous research has established an expected number of factors, analysts shift their focus from exploratory to confirmatory structural analysis.

Reliability

Reliability is an assessment of the consistency of a measurement, most commonly assessed with coefficient alpha (Cortina, 1993; Nunnally, 1978; Schmitt, 1996). Coefficient alpha can theoretically range from 0 (completely unreliable) to 1 (perfectly reliable), and only positive values are interpretable. Strictly speaking, coefficient alpha is a measure of internal consistency, but it is the most widely used method for estimating reliability, both for the overall measurement of a questionnaire and for any subscales supported by factor analysis. Despite some criticisms against its use (Sijtsma, 2009), it has a mathematical relationship to more direct estimates of reliability (e.g., test–retest) in that it provides a lower bound estimate of reliability. Thus, estimates of coefficient alpha provide a conservative estimate of reliability. Furthermore, there are well-established guidelines for acceptable values of coefficient alpha in the development of standardized questionnaires, with an acceptable range from 0.70 to 0.95 (Landauer, 1997; Lindgaard & Kirakowski, 2013; Nunnally, 1978).

Criterion-related validity

The assessment of criterion-related validity is the correlation between the measure of interest and a different measure, either one taken at the same time (concurrent validity) or later (predictive validity). High correlations between measurements believed to be related to the same construct are evidence of convergent validity. Low correlations between variables that are not expected to measure the same thing are evidence of divergent (discriminant) validity. A common minimum criterion for the absolute magnitude of correlations that support the hypothesis of convergent validity is 0.30 (Nunnally, 1978).

Norms

By itself, a score (individual or average) has no meaning. Meaning arises from comparison. When a metric is initially developed, it can be used to compare two or more groups using standard experimental designs (e.g., different products or different user groups). Over time, with the collection of sufficient data, it is possible to enhance the interpretation of scores through the development of norms (Sauro & Lewis, 2016).

Normative data are collected from one or more representative groups who have completed the questionnaire in a specified setting. Comparison with norms allows assessment of how good or bad a score is, within appropriate limits of generalization. With norms there is always a risk that the new sample does not match the normative group(s) (Anastasi, 1976), so it is important to understand where the norms came from when using them to interpret new scores. Development and maintenance of norms can be an expensive endeavor, so most questionnaires developed to assess perceived usability do not have them. Exceptions to this are the normative databases developed and maintained for the SUMI (Kirakowski & Corbett, 1993), Website Analysis and Measurement Inventory (Kirakowski & Cierlik, 1998), and Standardized User Experience Percentile Rank Questionnaire (SUPR-Q, Sauro, 2015), all of which require payment of a license fee but in return have professionally curated norms.

2. Early research: 1984–2007

2.1. Initial development and assessment of the SUS (Brooke, 1996)

Brooke (1996) first came up with 50 statements intended to address a range of possible user reactions to different aspects of system usability. He then had 20 people from his office systems engineering group with a variety of backgrounds (e.g., secretary, systems programmer) perform tasks with two software applications, one known to be relatively easy to use and the other quite difficult. The 10 items that become the standard SUS were those that were the most discriminating between the easy and difficult applications in this relatively small-sample experiment. Brooke noted that the absolute values of the correlations among the items were high (0.7–0.9).

Based on these findings, Brooke (1996) stated, "SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own." This original research seems to have had adequate content validity, but no metrics were provided for construct validity, concurrent validity, or reliability, very likely because the sample size of 20 was too small. Before the publication of Brooke (1996), however, Lucey (1991), in an unpublished thesis, reported that the reliability of the SUS (coefficient alpha) was an acceptable 0.85.

2.2. Research hiatus from 1997 through 2003

As far as I know, no research was published on the SUS during the 7 years from the publication of Brooke's short paper in 1996 through 2003.

2.3. Research resumes: Tullis and Stetson (2004) and Finstad (2006)

At the 2004 conference of the Usability Professionals Association, Tullis and Stetson presented a comparison of standardized usability questionnaires that included the SUS, QUIS, CSUQ, Words (ratings based Microsoft's Product Reaction Cards, Benedek & Miner, 2002), and an internally developed Fidelity questionnaire. A total of 123 Fidelity employees used a randomly assigned method to rate their experiences when completing two tasks with two websites. Across all methods, one site was significantly preferred over the other. In the most interesting result, *t*-tests of randomly selected subsamples of the data (with *n* of 6, 8, 10, 12, and 14) found that the SUS was the fastest to converge on the final (correct) conclusion regarding the preferred site. Where "agreement" means that the subsample *t*-test had a significant outcome that matched that of the full-sample *t*-test, the SUS reached 75% agreement at a sample size of 8 and 100% agreement when the sample size was 12. The CSUQ was the second fastest (75% agreement when $n = 8$ and 90% when $n = 12$). In contrast, even when $n = 14$, the other methods were in the low- to mid-70% of agreement with the full-sample decision. The key takeaway from this experiment was that user experience practitioners and researchers should seriously consider using the SUS as an efficient measure of perceived usability.

Finstad (2006) published a short research paper on the SUS in which he documented difficulty that non-native speakers had understanding the word "cumbersome" in the original version of Item 8 ("I found the system very cumbersome to use"). The key takeaway from this research was to use "awkward" in place of "cumbersome" in Item 8 (as shown in Figure 1).

3. Research since 2008

3.1. The evolution of norms for the SUS

The paper "An empirical evaluation of the System Usability Scale" (Bangor, Kortum, & Miller, 2008) was a seminal publication in the history of SUS research. As of 15th March 2018, according to Google Scholar, it has been cited over 1270 times. It is currently the most read paper in the history of the *International Journal of Human-Computer Interaction*, with over 7500 views. Its importance in driving the subsequent explosion of SUS research since its publication cannot be overstated.

Bangor et al. (2008) presented findings based on having used the SUS for almost 10 years in the evaluation of numerous products in various phases of development (based on more than 2300 completed SUS questionnaires collected over more than 200 studies). Their analyses and experience indicated that the SUS was a "highly robust and versatile tool for usability professionals" (p. 574). Some of their key findings were:

- The mean across all individual questionnaires was about 70, as was the mean computed across studies.
- Individual SUS scores ranged from 0 to 100, but across studies, the range of the means was more restricted, with 6% lower than a score of 50 and none lower than 30.
- Individual scores had a negative skew, but the distribution of study means was more normal.

- Inter-item correlations were consistently significant, ranging from 0.34 to 0.69.
- The SUS had an acceptable level of reliability (coefficient alpha of 0.91).
- The 10 items of the SUS all appeared to load on a single underlying factor.
- Comparison of six different classes of interface types (cell phones, customer equipment, graphical user interface, interactive voice response, Web, and Internet-based Web/IVR) found significant differences in SUS ratings as a function of interface type, which is evidence of scale sensitivity.
- There was evidence of a slight but significant negative relationship between score and age.
- There was no significant difference between male and female scores.
- Changes in SUS scores tracked logically with critical events in the product lifecycle process in a case study of iterative testing.

Given the large amount of SUS data collected over a decade, Bangor et al. (2008) made two attempts at developing norms with their data. About 10% (212) of the completed SUS questionnaires included an 11th item, an adjective rating scale with seven response options: 1: Worst imaginable (n = 1), 2: Awful (n = 0), 3: Poor (n = 15), 4: OK (n = 36), 5: Good (n = 90), 6: Excellent (n = 69), and 7: Best imaginable (n = 1). The SUS means for responses from 3 to 6 (for which $n \geq 15$) were, respectively after rounding to the nearest point, 39, 52, 73, and 86. The second approach was an absolute grading scale with A: 90–100, B: 80–89, C: 70–79, D: 60–69, and F: < 60.

Bangor, Kortum, and Miller (2009) increased the sample size of concurrent collection of SUS with the adjective rating scale to almost 1,000 cases. They reported a large and statistically significant correlation of 0.82 between the SUS and the adjective rating scale (evidence of concurrent validity). The means (and parenthetical sample sizes) for the seven response options were:

- 1: Worst imaginable = 12.5 (n = 4)
- 2: Awful = 20.3 (n = 22)
- 3: Poor = 35.7 (n = 72)
- 4: OK = 50.9 (n = 211)
- 5: Good = 71.4 (n = 345)
- 6: Excellent = 85.5 (n = 289)
- 7: Best imaginable = 90.9 (n = 16)

Note that Bangor et al. (2009) expressed some reservation over the interpretation of "OK" (with an associated mean SUS of 50.9) as suggesting an acceptable experience given an overall mean SUS closer to 70 in their large-sample data (Bangor et al., 2008). "In fact, some project team members have taken a score of OK to mean that the usability of the product is satisfactory and no improvements are needed, when scores within the OK range were clearly deficient in terms of perceived usability" (Bangor et al., 2009, p. 120). Their current practice is to anchor this response option with "Fair" instead of "OK" (Phil Kortum, personal communication, 22nd February 2018).

This line of research inspired the development of a curved rather than an absolute grading scale for the SUS (Sauro, 2011; Sauro & Lewis, 2012, 2016). Bangor et al. generously shared their SUS data with Jeff Sauro, as did Tullis and Albert (2008). With this combined data set from 446 studies and over 5000 individual SUS responses, Sauro (2011) used a logarithmic transformation on reflected scores to normalize the distribution, then computed percentile ranks for the entire range of SUS scores. Sauro and Lewis (2012, 2016)) used those percentile ranks to create the curved grading scale (CGS) shown in Table 1.

Note that the average score in the data used to create the Sauro–Lewis CGS was 68, which was by design the exact center of the CGS (a grade of C), but would have been a D in the absolute grading scale. With its 11 grade ranges, the CGS also provides a finer-grained scale than the adjective scale with its seven response options. It addresses the weakness of "OK" in the adjective scale because a 50 would receive an F (clearly deficient) while the lowest value in the range for C (an average experience) is 65. Finally, the CGS is consistent with an industrial practice that has become increasingly common of interpreting a mean SUS of at least 80 (A−) as indicative of an above-average user experience. Throughout the rest of this article, letter grades are from the Sauro–Lewis CGS.

3.2. Additional normative research

The Sauro–Lewis CGS provides good general guidance for the interpretation of SUS means. Several lines of research have shown, however, that different types of products and interfaces differ significantly in perceived usability. For example, Sauro (2011) partitioned his data from 446 studies into groups based on product type. The means (with associated CGS grades and number of studies) for some of the key categories were:

- Business-to-business software: 67.6 (C, n = 30)
- Mass market consumer software: 74.0 (B−, n = 19)
- Public facing websites: 67.0 (C, n = 174)
- Internal productivity software: 76.7 (B, n = 21)

Kortum and Bangor (2013) published SUS ratings of overall experience for a set of 14 everyday products from a survey of more than 1000 users. Examples of the SUS means (with associated CGS grades and number of respondents) for products with low, medium, and high perceived usability were:

- Excel: 56.5 (D, n = 866)
- Word: 76.2 (B, n = 968)
- Amazon: 81.8 (A, n = 801)
- Google search: 92.7 (A+, n = 948)

Table 1. The Sauro–Lewis CGS.

SUS Score range	Grade	Percentile range
84.1–100	A+	96–100
80.8–84.0	A	90–95
78.9–80.7	A−	85–89
77.2–78.8	B+	80–84
74.1–77.1	B	70–79
72.6–74.0	B−	65–69
71.1–72.5	C+	60–64
65.0–71.0	C	41–59
62.7–64.9	C−	35–40
51.7–62.6	D	15–34
0.0–51.6	F	0–14

In 2015, Kortum and Sorber collected SUS ratings from 3,575 users on the usability of 15 mobile applications for phones and tablets (10 based on popularity and 5 that users identified as using frequently). The mean SUS for the top 10 applications was 77.7 (a B+) with a difference of about 20 points between the highest- (87.4, A+) and lowest- (67.7, C) rated applications. This range from grades of C to A+ is skewed to the high end of the scale, but this is likely due to the method used to select the applications for the study (high popularity and high frequency of use).

There are different ways to interpret what these findings (Kortum & Bangor, 2013; Kortum & Sorber, 2015; Sauro, 2011) mean for industrial practice in user experience engineering. They could be interpreted as diminishing the value of the more general norms embodied in the CGS, but a more pragmatic interpretation is that they enhance the general norms. For example, consider the Kortum and Bangor ratings of everyday products. It should not be surprising that a complex spreadsheet program has lower perceived usability than a well-designed search box. For many projects, setting a SUS benchmark of 80 (A−) is reasonable and achievable. If, however, the project is to develop a competitive spreadsheet application, a SUS of 80 is probably unrealistically high (and is probably unrealistically low if developing a new search interface). When possible, practitioners should use a combination of comparison with norms and competitive evaluation when assessing the quality of their products. Practitioners should also exercise some caution when using data from within-subjects studies as benchmarks because respondents who are comparing products may, to a currently unknown extent, give slightly lower ratings to harder products and higher ratings to easier products than they otherwise might.

Brooke (1996) cautioned against attempts to extract meaning from the items of the SUS, specifically, that the "SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own" (p. 189). At the time, this admonition was appropriate because his analyses were based on data from 20 people. With substantially more data in hand, Lewis & Sauro, In press developed a series of regression equations for setting benchmarks for SUS items. This would not be useful for practitioners who are collecting data for attributes that are not one of the SUS items, such as findability. There are, however, some SUS items that might sometimes be useful apart from their contribution to the overall SUS, in particular Item 2 (perceived complexity), Item 3 (perceived ease-of-use), Item 6 (perceived consistency), Item 7 or 10 (perceived learnability), and Item 9 (confidence in use). The data used to develop the regression equations came from 166 unpublished usability studies/surveys (a total of 11,855 individual SUS questionnaires). The 10 regression equations (with the text of the associated SUS item), computed using the means from the 166 individual studies, were:

- SUS01 = 1.073927 + 0.034024(SUS): "I think that I would like to use this system frequently."
- SUS02 = 5.834913 – 0.04980485(SUS): "I found the system unnecessarily complex."
- SUS03 = 0.4421485 + 0.04753406(SUS): "I thought the system was easy to use."
- SUS04 = 3.766087 – 0.02816776(SUS): "I think that I would need the support of a technical person to be able to use this system."
- SUS05 = 1.18663 + 0.03470129(SUS): "I found the various functions in this system were well integrated."
- SUS06 = 4.589912 – 0.03519522(SUS): "I thought there was too much inconsistency in this system."
- SUS07 = 0.9706981 + 0.04027653(SUS): "I would imagine that most people would learn to use this system very quickly."
- SUS08 = 5.575382 – 0.04896754(SUS): "I found the system very awkward to use."
- SUS09 = 0.6992487 + 0.04435754(SUS): "I felt very confident using the system."
- SUS10 = 4.603949 – 0.03692307(SUS): "I needed to learn a lot of things before I could get going with this system."

Note that due to the mixed tone of the SUS items the directionality of benchmarks would be different for odd- and even-numbered items. For odd-numbered items, higher scores are better (using the basic five-point item scale shown in Figure 1); for even-numbered items lower scores indicate a better user experience. The first step in using the equations is to select a SUS value corresponding to a desired CGS grade level. For example, if a practitioner is interested in interpreting Item 3, "I thought the system was easy to use," then a mean score of 3.67 would correspond to a SUS mean of 68 (an average overall system score). For consistency with an above-average SUS mean of 80, the corresponding target for Item 3 would be an average score of at least 4.24 (ideally statistically greater than the benchmark to control the risk of exceeding it by chance).

3.3. The factor structure of the SUS

The SUS was designed to produce a single overall measure of perceived usability (Bangor et al., 2008; Brooke, 1996). In 2009, factor analyses of three large-sample studies consistently indicated that Items 4 and 10 aligned on their own factor, separate from the other eight items. First, Lewis and Sauro (2009) reanalyzed the inter-item correlation matrix published by Bangor et al. (2008) and an independent set of data (324 SUS questionnaires collected over 19 studies), finding that their two-factor patterns matched with Items 4 and 10 on one factor and the remaining items on the other. Based on the item content, Lewis and Sauro named the subscales associated with these factors Learnable (Items 4 and 10) and Usable (the remaining items). Later that year, Borsci, Federici, and Lauriola (2009) reported an independent replication of that finding using an Italian version of the SUS and a different analytical method. The promise of this research was that practitioners could, with little additional effort, extract more information from their SUS data.

Unfortunately, factor analyses conducted since 2009 (Kortum & Sorber, 2015; Lewis, Brown, & Mayes, 2015; Lewis, Utesch, & Maher, 2013, 2015; Sauro & Lewis, 2011) failed to replicate the two factors that seemed apparent in 2009. The results of Borsci, Federici, Gnaldi, Bacci, and

Bartolucci (2015) suggested the possibility that the alignment of SUS items on Usable and Learnable subscales might depend on the level of user experience with the rated system, but Lewis, Utesch, & Maher (2015) did not replicate this finding. Aside from Borsci, Federici, Gnaldi et al. (2015), the analyses since 2009 have been somewhat consistent with the alignment of positive- and negative-tone items on separate factors. This is a type of unintentional factor structure reported to occur with sets of mixed-tone items (Barnette, 2000; Davis, 1989; Pilotte & Gable, 1990; Schmitt & Stuits, 1985; Schriesheim & Hill, 1981; Stewart & Frye, 2004; Wong, Rindfleisch, & Burroughs, 2003).

Lewis and Sauro (2017b) assembled a data set of 9,156 completed SUS questionnaires from 112 unpublished industrial usability studies and surveys for a range of software products and websites. Both exploratory and confirmatory factor analyses of the ratings were consistent with a two-factor structure driven by the tone of the items. Lewis and Sauro (p. 7) concluded:

> It is possible that the SUS might have internal structure that is obscured by the effect of having mixed tone items, but we found no significant evidence supporting that hypothesis. It is interesting to note … that the magnitude of the factor loadings for Items 4 and 10 in all three exploratory analyses were greater than those for Items 2, 6, and 8 on the negative tone factor, suggesting (but not proving) that there might be some research contexts in which they would emerge as an independent factor. Because a distinction based on item tone is of little practical or theoretical interest when measuring with the SUS, it is, with some regret but based on accumulating evidence, that we recommend that user experience practitioners and researchers treat the SUS as a unidimensional measure of perceived usability, and no longer routinely compute or report Usability and Learnability subscales.

3.4. Other psychometric findings

Table 2 summarizes the findings since 2008 regarding key psychometric properties of the SUS, limited to research conducted with the standard English version (as shown in Figure 1).

For this body of research, estimates of reliability using coefficient alpha ranged from 0.83 to 0.97, averaging around 0.91. The sample sizes had considerable variability, from dozens to hundreds to thousands of cases depending on the study.

Estimates of concurrent validity had significant correlations ranging from 0.22 to 0.96. Typically, estimates of subjective usability tend to correlate more highly with each other than with estimates of objective usability (Sauro & Lewis, 2009). That pattern generally held in Table 2, with correlations of the SUS with other subjective ratings of usability such as user friendliness, Usability Metric for User Experience (UMUX, Finstad, 2010), UMUX-LITE (a shorter version of the UMUX, Lewis et al., 2013), the SUPR-Q (Sauro, 2015), adjective rating scale (Bangor et al., 2008, 2009), and ratings of likelihood-to-recommend ranging from 0.50 to 0.96. Correlations with the objective metric of successful task completions (success rate) ranged from 0.22 to 0.50. There was evidence of significant sensitivity of the SUS to independent variables such as different products/systems, changes to products, amount of experience with a product/system, personality type, application type, and platform type (phone vs. tablet).

Taken together, these findings indicate that the SUS has excellent reliability and concurrent validity with other measures of perceived and objective usability, which lead to sensitivity when comparing perceived usability as a function of a variety of independent variables (products/systems, mobile apps, mobile platforms, and personality types). Of particular interest to researchers and practitioners is the robust sensitivity to the amount (duration and/or frequency) of experience users have with the products and systems they rated. When making comparisons with the SUS (as with any other usability metric such as successful task completions or task completion times), it is important to track and control for differences in the amount of user experience.

Table 2. Additional psychometric findings for the standard SUS.

Study	Sample size	Reliability (Coefficient alpha)	Concurrent validity (correlation)	Validity details	Evidence of sensitivity
Bangor et al. (2008)	2324	0.91	0.81	Rating of user friendliness	Product differences, changes to products
Bangor, Joseph, Sweeney-Dillon, Stettler, and Pratt (2013)	872	NA	NA	NA	Prediction of business indicators
Berkman and Karahoca (2016)	151	0.83	0.74	UMUX	NA
Finstad (2010)	558	0.97	0.96	UMUX	System differences
Kortum and Bangor (2013)	1058	NA	0.50–0.79	Adjective rating scale	Product differences, amount of experience
Kortum and Johnson (2013)	31	NA	NA	NA	Amount of experience
Kortum and Oswald (2017)	268	NA	NA	NA	Personality type
Kortum and Sorber (2015)	3575	0.88	NA	NA	App type, platform type, operating system, amount of experience
Lah and Lewis (2016)	18	NA	0.50	Success rate	Amount of experience
Lewis (2018)	618	0.93	0.76, 0.79, 0.74	CSUQ, UMUX, UMUX-LITE	Operating system
Lewis et al. (2015)	471	0.90	0.50, 0.63	Success rate, likelihood-to-recommend	NA
Lewis and Sauro (2009)	324	0.92	NA	NA	Product differences
Lewis and Sauro (2017b)	9156	0.91	NA	NA	NA
Lewis et al. (2013)	389	0.89	0.90, 0.81	UMUX, UMUX-LITE	NA
McLellan, Muddimer, and Peres (2012)	262	NA	NA	NA	Amount of experience
Peres, Pham, and Phillips (2013)	85	NA	0.22	Success rate	NA
Sauro (2015)	3891	NA	0.75	SUPR-Q	NA
Sauro and Lewis (2011)	107	0.92	NA	NA	NA

3.5. The flexibility of the SUS

Another of the strengths of the SUS in practical UX work is its flexibility, which extends beyond minor wording changes. In recent years, researchers have expanded its use beyond traditional usability testing to the retrospective measurement of perceived usability of products or classes of products (Grier, Bangor, Kortum, & Peres, 2013; Kortum & Bangor, 2013). Grier (2013) described a version of the SUS altered for the context acquiring products for the U.S. military that are easy to troubleshoot and maintain, but did not provide any assessment of its psychometric properties.

Positive version of the SUS

Sauro and Lewis (2011) explored a more extreme manipulation of the SUS, specifically, changing the tone of the even-numbered items from negative to positive, as shown in Figure 2.

The original SUS was designed in accordance with a common strategy to control acquiescence bias, the hypothesized tendency of respondents to agree with statements, by having respondents rate statements with a mix of positive and negative tone. This practice also has potential benefits in helping researchers identify respondents who were not attentive to the statements they rated. There is, however, evidence that including a mix of positively and negatively worded items can create more problems than it solves (Barnette, 2000; Stewart & Frye, 2004), lowering internal reliability, distorting factor structure, and increasing interpretation problems in cross-cultural research. Furthermore, respondents may have difficulty switching response behaviors when completing questionnaires with mixed-tone items (mistakes), and researchers might forget the necessary step of reversing item scores for negative-tone items when computing overall scores (miscoding).

Sauro and Lewis (2011) administered a retrospective survey using the standard and positive versions of the SUS (n = 213) across seven websites. The reliability (coefficient alpha) of both questionnaires was high (Standard: 0.92; Positive: 0.96). The mean SUS scores for the two versions were not significantly different (Standard: 52.2; Positive: 49.3; $t(206) = 0.85$, $p > 0.39$). They found no evidence of acquiescence bias in either version, but estimated that about 17% of the completed standard questionnaires contained mistakes. They also reported that three of 27 SUS data sets (11%) contributed by anonymous donors to additional research efforts had miscoded SUS scores. "The data presented here suggest the problem of users making mistakes and researchers miscoding questionnaires is both real and much more detrimental than response biases" (Sauro & Lewis, 2011, p. 2221).

	The System Usability Scale Positive Version		Strongly Disagree 1	2	3	4	Strongly Agree 5
1	I think that I would like to use the website frequently.		o	o	o	o	o
2	I found the website to be simple.		o	o	o	o	o
3	I thought the website was easy to use.		o	o	o	o	o
4	I think that I could use the website without the support of a technical person.		o	o	o	o	o
5	I found the various functions in the website were well integrated.		o	o	o	o	o
6	I thought there was a lot of consistency in the website.		o	o	o	o	o
7	I would imagine that most people would learn to use the website very quickly.		o	o	o	o	o
8	I found the website very intuitive.		o	o	o	o	o
9	I felt very confident using the website.		o	o	o	o	o
10	I could use the website without having to learn anything new.		o	o	o	o	o

Figure 2. The positive SUS.

Additional research using the positive version of the SUS has provided evidence of its reliability, validity, and sensitivity. In Lewis et al. (2013, n = 402) its estimated reliability was 0.94, and in follow-on work (Lewis, Utesch, & Maher, 2015, n = 397) it was 0.91. It correlated significantly with concurrently collected UMUX (0.79) and UMUX-LITE (0.85) scores (Lewis et al., 2013) as well as ratings of overall experience (0.67) and likelihood-to-recommend (0.71) (Lewis, Utesch, & Maher, 2015). It has also been found to be sensitive to the amount of product experience and self-reported levels of product expertise (Lewis, Utesch, & Maher, 2015).

As flexible as it is, it is possible to break the SUS with items rewritten to be extreme. Sauro (2010) described an experiment in which SUS items were manipulated to investigate two variables: item intensity and item tone. For example, the extreme negative version of the SUS Item 4 was "I think that I would need a permanent hot-line to the help desk to be able to use the website." The 62 participants were volunteers attending the 2008 conference of the Usability Professionals Association (UPA). They used one of five questionnaires to rate the UPA website: all positive extreme, all negative extreme, mixed extreme version 1, mixed extreme version 2, or the standard SUS. The scores from the all positive extreme and all negative extreme were significantly different from the standard SUS.

> By rephrasing items to extremes, only respondents who passionately favored the usability of the UPA website tended to agree with the extremely phrased positive statements, resulting in a significantly lower average score. Likewise, only respondents who passionately disfavored the usability agreed with the extremely negative statements, resulting in a higher average score. (Sauro & Lewis, 2016, p. 211).

Can i leave this one out?

To compute the overall SUS score, respondents must provide a rating for each item. The instruction that Brooke (1996, p. 193) provided in the initial publication of the SUS was, "All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale." Thus, the typical practice when respondents do not provide a rating for an item is to replace the blank with the default rating of 3.

But what if there is an item that would be confusing or distracting to respondents in a particular context of measurement? For example, the first SUS item is "I think I would like to use this system frequently." If the system under study is one that would only be used infrequently (e.g., a troubleshooting process or system for registering complaints), then there is a concern that including this item would distort the scores, or at best, distract the participant.

Lewis and Sauro (2017a) investigated the consequences of removing individual items from the standard SUS. Because previous research had indicated that small amounts of data missing from standardized usability questionnaires had little

effect on the resulting scores (Lah & Lewis, 2016; Lewis, 2002) and the items of the SUS are significantly intercorrelated (Brooke, 1996), they hypothesized that the 10 possible nine-item versions of the SUS should not differ much from the score obtained with all 10 items given appropriate adjustment of the of the SUS multiplier.

To understand how to adjust the SUS multiplier, consider how the standard multiplier works. The process of determining score contributions described in the introduction results in a score that, without multiplication, would range from 0 to 40 (a maximum score contribution of 4 multiplied by 10 items). To stretch that out so it ranges from 0 to 100, it is necessary to multiply the sum of the score contributions by 100/40, which is the derivation of the "2.5" multiplier. After dropping one item, the score contributions can range from 0 to 36 (9 × 4). To stretch this out to range from 0 to 100, the multiplier needs to be 100/36.

Lewis and Sauro (2017a) analyzed a data set of 9,156 completed SUS questionnaires from 112 unpublished industrial usability studies and surveys. Note that with n = 9,156, the study had the power to reliably detect very small differences and to precisely compute confidence intervals around estimated means, allowing a focus on differences that have practical rather than simply statistical significance (which only supports claims that differences are not plausibly 0). They computed the 10 possible nine-item scores that are possible when leaving one SUS item out, following the standard scheme for computing these SUS scores but multiplying the sum of the score contributions by 100/36 instead of 2.5 to compensate for the missing item. For each nine-item variant of the SUS, they assessed scale reliability using coefficient alpha, the correlation with the standard SUS, and the magnitude of the mean difference.

As expected, all nine-item variants of the SUS correlated significantly with the standard SUS (all $r > 0.99$). Dropping one item had no appreciable effect on scale reliability, with all values of coefficient alpha ranging from 0.90 to 0.91. The mean scores of all 10 possible nine-item variants of the SUS were within one point (out of a100) of the mean of the standard SUS. Thus, it appears that practitioners can leave out any one of the SUS items without having a practically significant effect on the resulting scores, as long as an appropriate adjustment is made to the multiplier (specifically, multiply the sum of the adjusted item scores by 100/36 instead of the standard 100/40, or 2.5, to compensate for the dropped item).

Research implications

Taking all of this research into account, the major conclusions are:

- The SUS can be used as a measure of perceived usability in standard task-based usability testing or as a retrospective measure in surveys.
- The purported advantages of including negative and positive items in usability questionnaires do not appear to outweigh the disadvantages.
- Researchers who use the standard SUS do not need to switch to the positive version, but do need to verify proper scoring.
- Practitioners who use the standard SUS in moderated usability testing should include procedural steps (e.g., during debriefing) to ensure error-free completion.
- It is more difficult to correct mistakes respondents make in retrospective surveys or unmoderated usability testing although in these large-sample methods such errors are unlikely to have a major impact on overall SUS scores.
- Researchers and practitioners who do not have a current investment in the standard SUS should consider using the positive version to reduce the likelihood of response or scoring errors, especially when conducting surveys or unmoderated remote usability studies.
- Until more data are published that compare the magnitudes of standard and positive versions of the SUS, researchers and practitioners should consider collecting data with both to verify correspondence in their domains of interest and, when possible, publishing those results.
- Practitioners who have a good reason to drop one item from the SUS can do so without appreciably affecting the resulting overall SUS score as long as they make the appropriate adjustment to the formula used to compute that score.

3.6. Translations of the SUS

Since 2014 there have been a number of published translations of the SUS, including Arabic, Slovene, Polish, Italian, Persian, and Portuguese. Table 3 summarizes the basic psychometric findings from that body of research.

The average estimate of reliability across these studies was about 0.81, lower than that typically found for the English version but well above the typical minimum criterion of 0.70. Estimates of concurrent validity with a variety of other metrics of perceived usability were significant correlations ranging from 0.45 to 0.95. Several studies found the SUS sensitive to the amount of experience with the product or system under investigation, consistent with sensitivity findings reported for the English version.

In Blažica and Lewis (2015), respondents rated the usability of the Slovene version of Gmail, providing an opportunity to compare those ratings with the English assessment of Gmail reported in Kortum and Bangor (2013). The overall mean from the Slovene version was 81.7 (n = 182, SD = 13.5, 95% confidence interval ranging from 79.7 to 83.7). This was close to the mean SUS of 83.5 for Gmail reported by Kortum and Bangor (2013) (n = 605, SD = 15.9, 95% confidence interval ranging from 82.2 to 84.8). These confidence intervals overlapped substantially, indicating that the Gmail results for the Slovene version were reasonably close to the value published by Kortum and Bangor.

Although there are no data currently available regarding their psychometric properties, versions of the SUS in German (Rummel, 2015) and Swedish (Göransson, 2011) are available. There is a clear need for translation into additional languages with proper assessment of psychometric properties and, when possible, studies of correspondence with published English norms (e.g., Kortum & Bangor, 2013).

Table 3. Psychometric findings for various SUS translations.

Study	Sample size	Language	Reliability (Coefficient alpha)	Concurrent validity (correlation)	Validity details	Evidence of sensitivity
AlGhannam et al. (2017)	90	Arabic	0.82	NA	NA	Amount of experience
Blažica and Lewis (2015)	182	Slovene	0.81	0.52	Likelihood-to-recommend	Amount of experience
Borkowska and Jach (2016)	NA	Polish	0.81	0.82	CSUQ	NA
Borsci, Federici, Gnaldi, Bacci, and Bartolucci (2015)	186	Italian	NA	0.55, 0.72	UMUX, UMUX-LITE	Amount of experience
Borsci, Federici, Gnaldi et al. (2015)	93	Italian	NA	0.45, 0.66	UMUX, UMUX-LITE	Amount of experience
Borsci et al. (2009)	196	Italian	0.81	NA	NA	NA
Borsci, Federici, Mele, and Conti (2015)	20	Italian	0.84	0.82–0.95	UMUX, UMUX-LITE	Sighted vs. blind users
Dianat et al. (2014)	202	Persian	0.79	NA	NA	NA
Martinsa, Rosa, Queirós, and Silva (2015)	32	Portuguese	NA	0.70	PSSUQ	NA

3.7. Correspondence with other metrics of perceived usability

Since Finstad (2010) published his measure of perceived usability, the UMUX, and found it to correspond closely to concurrently collected SUS scores, several other researchers have published similar findings with the UMUX, UMUX-LITE, UMUX-LITEr, and CSUQ. Just because metrics correlate significantly, they do not necessarily have similar magnitudes when placed on a common scale. Table 4 provides a summary of the findings from six studies of the differences in concurrently collected SUS means and other metrics of perceived usability. Three of the studies (Borsci, Federici, Gnaldi et al., 2015; Finstad, 2010; and Lewis et al., 2013) had data from two independent surveys, so each of those studies have data in two rows of the table, for a total of nine estimates of correspondence. For some studies the value of the mean SUS was confidential, but the difference scores are available for publication and, for this analysis, it is the differences that matter.

For detailed information about the other metrics of perceived usability, see Lewis (2018) or the original sources (UMUX: Finstad, 2010; UMUX-LITE and UMUX-LITEr: Lewis et al., 2013; CSUQ: Lewis, 1995). Following are short descriptions.

UMUX

The UMUX was designed to get a measurement of perceived usability consistent with the SUS but using fewer items, just four, that closely conformed to the ISO (1998) definition of usability. The four UMUX items use a seven-point scale and, like the standard SUS, are a mix of positive and negative tone.

UMUX-LITE

The UMUX-LITE is a two-item questionnaire made up of the positive-tone items from the UMUX. This resulted in a parsimonious questionnaire that has a connection through its items to Davis' (1989) Technology Acceptance Model (TAM), which assesses the dimensions of usefulness and ease-of-use.

UMUX-LITEr

The UMUX-LITEr is a regression-adjusted version of the UMUX-LITE (UMUX-LITEr = 0.65(UMUX-LITE) + 22.9). Lewis et al. (2013) developed the adjustment to compensate for a small but statistically significant difference they observed between concurrently collected SUS and UMUX-LITE data. A practical consequence of applying the regression equation is that UMUX-LITEr scores can only range from 22.9 to 87.9 rather than 0 to 100. This is not a serious problem for comparison with SUS means, however, because although individual SUS scores can and do range from 0 to 100, SUS means are almost always greater than 0 and less than 100 (Kortum & Acemyan, 2013).

CSUQ

The CSUQ is a variant of the PSSUQ, modified from it for use as a retrospective measure of perceived usability. The current version (Sauro & Lewis, 2016) has 16 items, all positive tone with seven-point scales in which lower numbers indicate a more favorable rating. Traditionally, the overall CSUQ score is the average of its 16 items, so it can take a value between 1 (best experience) and 7 (worst experience). To convert it to a 0–100 point scale in which larger numbers indicate a better experience, use the following equation (Lewis, 2018): CSUQ = 100 − (((CSUQ01 +

Table 4. Differences between SUS and other measures of perceived usability.

Study	Sample size	UMUX	UMUX-LITE	UMUX-LITEr	CSUQ	SUS Version
Berkman and Karahoca (2016)	151	1.3	1.0	5.5	−0.5	Standard English
Borsci et al. (2015)	186	−13.8	NA	−2.9	NA	Standard Italian
Borsci, Federici, Gnaldi et al. (2015)	93	−12.5	NA	−1.3	NA	Standard Italian
Finstad (2010): Study 1	273	1.1	NA	NA	NA	Standard English
Finstad (2010): Study 2	285	0.5	NA	NA	NA	Standard English
Lewis (2018)	618	−2.4	−4.0	−2.2	−2.0	Standard English
Lewis et al. (2013): Study 1	389	0.4	3.2	−0.7	NA	Standard English
Lewis et al. (2013): Study 2	402	2.7	3.7	0.1	NA	Positive English
Lewis, Utesch, & Maher (2015)	397	NA	5.7	1.2	NA	Positive English
Mean (Overall)	310	−2.8	1.9	0.0	−1.3	All versions
Mean (English)	359	0.6	1.9	0.8	−1.3	English
Mean (Standard English)	343	0.2	0.1	0.9	−1.3	Standard English

CSUQ02 + CSUQ03 + CSUQ04 + CSUQ05 + CSUQ06 + CSUQ07 + CSUQ08 + CSUQ09 + CSUQ10 + CSUQ11 + CSUQ12 + CSUQ13 + CSUQ14 + CSUQ15 + CSUQ16)/16) - 1)(100/6). Breaking this down, the process of getting from a traditional CSUQ score to one that matches the SUS involves subtracting 1 from the mean of the 16 individual CSUQ items and multiplying that by 100/6 to stretch it out to a 0–100 point scale, then subtracting from 100 to reverse the scale.

Analysis of correspondence

As shown in Table 4, averaging across all nine estimates of differences with concurrently collected SUS means, the UMUX-LITEr had the closest correspondence with the SUS – a mean difference of 0. The UMUX had the largest mean difference, just under 3 points out of the 0–100-point scale. The largest discrepancies in the table were those reported by Borsci, Federici, Gnaldi et al. (2015) for the UMUX. Without the Borsci, Federici, Gnaldi et al. (2015) findings, which used an Italian version of the SUS and UMUX, both the UMUX and UMUX-LITEr were less than 1 point different from the matching SUS means. Removing the means from data sets that used the positive version of the SUS led to all UMUX-related metrics being less than 1 point different from matching SUS means. The CSUQ has been compared to the SUS in only two studies, both of which used the standard version of the SUS, with a mean difference from matching SUS means of just over 1 point.

Although there are still gaps to fill with future research, it appears that the correspondence between the SUS and these alternate measures of perceived usability is very close. This finding is encouraging for researchers and practitioners who use any of these methods for measuring perceived usability because not only do they appear to measure the same underlying construct, they also appear to produce scores that, when placed on a common 101-point scale (0–100), have very similar magnitudes. A practical consequence of this correspondence is that practitioners who use one of these other metrics of perceived usability can, with some confidence but appropriate caution, use SUS norms like the Sauro–Lewis CGS to interpret their means. On the other hand, because there are only a few published studies in which the SUS and UMUX-related measures have been concurrently collected, practitioners should exercise caution in switching from the SUS to a UMUX-related metric without checking for concurrence in their research or practice context.

3.8. The relationship between the SUS, likelihood-to-recommend, and the net promoter score

Since its introduction to the business world in 2003 (Reichheld, 2003, 2006), the net promoter score (NPS) has become a popular metric of customer loyalty in industry. The NPS uses a single likelihood-to-recommend question ("How likely is it that you would recommend our company to a friend or colleague?") with 11 scale steps from 0 (Not at all likely) to 10 (Extremely likely). In NPS terminology, respondents who select a 9 or 10 are "Promoters," those selecting 0 through 6 are "Detractors," and all others are "Passives," with the NPS computed as the percentage of Promoters minus the percentage of Detractors.

Investigation of the relationship between SUS and the likelihood-to-recommend rating that underlies the NPS has consistently shown significant correlation. Regression analyses of concurrently collected SUS and likelihood-to-recommend data from 2201 users and over 80 products found a strong positive correlation of 0.623, meaning SUS scores explained about 39% of the variability in responses to the likelihood-to-recommend question (Sauro & Lewis, 2016). This leaves 61% of the variability unexplained, so although SUS and likelihood-to-recommend have a strong relationship, there appear to be factors other than perceived usability that affect loyalty as measured with likelihood-to-recommend and its derived metric, the NPS.

3.9. Other miscellaneous research findings

It is common in market research to account for differences in gender, age, and geographic location. The published findings are somewhat mixed with regard to the influence of gender on SUS ratings. Of six studies that investigated the effect of gender, five found no significant effect (Bangor et al., 2008; Berkman & Karahoca, 2016; Kortum & Bangor, 2013; Kortum & Sorber, 2015; Tossell, Kortum, Shepard, Rahmati, & Zhong, 2012), one found a statistically significant difference. One found a significant difference but reported that the apparent gender difference was likely due to differences in personality characteristics (Kortum & Oswald, 2017). Two studies examined the effect of age on SUS scores, with both reporting no significant difference (Bangor et al., 2008; Berkman & Karahoca, 2016).

Regarding effects of geography, Kortum and Acemyan (2018) collected SUS ratings of 11 popular products from 3,168 residents of the United States recruited through Amazon Mechanical Turk. Analysis of results as a function of geographic region (based on participant self-report of state of residence paired with US Census Bureau geographic and population density divisions) found little variation in SUS means. For the nine US Census Bureau geographic divisions, the mean SUS ranged from 73.1 (B–) to 74.7 (B), a nonsignificant difference of 1.6 points crossing a single grade boundary of the Sauro–Lewis CGS. Differences as a function of population density (rural, urban cluster, and urban) were also nonsignificant, even with the very large-sample size.

Kortum and Oswald (2017) investigated the impact of personality on SUS scores. People's scores on personality traits have been shown to be reliable and predict important outcomes in work, school, and life domains. In this study, 268 participants used the SUS to retrospectively assess the perceived usability of 20 different products. Participants also completed a personality inventory which provides measurement of five broad personality traits: Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. They found significant correlation between the SUS and measures of Openness to Experience and Agreeableness.

Of these four independent variables, gender, age, geography, and personality traits, the only variable for which there is

compelling evidence of having an effect on SUS scores is personality. Although the existing research indicates little or no effect of gender or age, it is reasonable for researchers and practitioners to include in their studies, as appropriate, a range of genders and ages, and to analyze and publish the effect of these variables on their SUS scores to increase the number of studies investigating these variables.

It is important for researchers and practitioners to be aware of potential effects of personality traits on the assessment of perceived usability, and to ensure that sample selection procedures, as appropriate, are unbiased with regard to the traits of Openness to Experience and Agreeableness. It seems unlikely that practitioners would routinely require participants to complete a personality inventory. It is important, however, for researchers to include this step to replicate and extend the work of Kortum and Oswald (2017), furthering our understanding of these effects.

4. The future of the SUS

The future of the SUS appears to be very bright for both research and practice in usability and user experience. Given its age one might expect interest in it to have waned, but instead the pace of research on the SUS is accelerating. Furthermore, the body of SUS research indicates that it is a powerful instrument for the assessment of perceived usability. That said, there is still substantial opportunity for additional research to fill in existing gaps.

4.1. Periodic refreshes and extensions of norms

The Sauro–Lewis CGS for the SUS (Sauro & Lewis, 2016) and research on everyday products by Kortum and his associates (Kortum & Acemyan, 2018; Kortum & Bangor, 2013; Kortum & Oswald, 2017) provide the best "open source" norms for any current license-free questionnaire that assesses perceived usability. An open question regarding these norms is the extent to which they might change over time and need to be refreshed.

The information used to produce the current Sauro–Lewis grading scale was all the data available circa 2011, which had a mean of 68. In the recent database assembled by Lewis & Sauro, In press, the overall SUS mean was 70.8. This is still a C on the CGS, but given that it includes more recent studies (n = 11,855 completed SUS questionnaires collected over 166 studies), this raises a question regarding whether or not there has been any shift in the relationship between an acceptable level of perceived usability and associated SUS scores. Additional topics for future normative research include investigation of other factors that might influence SUS scores such as method of data collection (e.g., survey vs. usability study; random survey vs. interrupt survey), effect of task difficulty, effect of experience with a narrow or broad range of competitive products?

Not only are these important topics for future research, it raises questions about the best way to gather new information and conduct periodic refreshes and extensions. Ideally, there would be an international clearinghouse for storing SUS data, gathered from published work and from donation of anonymized industrial studies, available to all interested researchers (perhaps with additional benefits for contributing researchers). The best candidates for managing such a clearinghouse would be a university or consortium of universities, or perhaps a professional international organization such as the ACM Special Interest Group for Computer–Human Interaction or the User Experience Professionals Association.

4.2. Additional translations

There is a clear need for translations of the SUS into more languages. In addition to following the steps that assure translation that is as accurate as possible, researchers who translate the SUS also need to conduct the appropriate psychometric research to establish the translation's reliability, validity, and sensitivity. Ideally, that research would include ratings of products that would allow the assessment of the correspondence between SUS scores from the translated and the standard English version (e.g., Blažica & Lewis, 2015).

4.3. Correspondence with additional metrics

The correspondence to date between the SUS, CSUQ, and UMUX-related metrics is encouraging, but additional studies would be helpful in providing a more solid foundation for claims of correspondence and understanding if there are any contexts of measurement in which the metrics systematically fail to correspond. It would also enhance our understanding of the relationship of the SUS with additional metrics. In particular, there are research gaps with regard to its relationship with the TAM (Davis, 1989) and metrics that attempt to assess the broader concept of user experience, of which perceived usability is one component (e.g., AttracDiff and the User Experience Questionnaire, Diefenbach, Kolb, & Hassenzahl, 2014; SUPR-Q, Sauro, 2015).

Researchers can use the SUS along with alternative metrics in regression analysis to assess the difference in explained variance of outcome metrics for models with and without the SUS. For example, Lewis and Mayes (2014) developed the Emotional Metric Outcomes (EMO) questionnaire to assess the emotional outcomes of interaction, especially the interaction of customers with service-provider personnel or software. The primary purpose of the EMO was to move beyond traditional assessment of satisfaction to achieve a more effective measurement of customers' emotional responses to products and processes. Concurrent measurement with the SUS indicated that the correlation of the SUS with likelihood-to-recommend ratings may primarily be due to emotional rather than utilitarian aspects of the SUS. They arrived at this conclusion because, when modeled alone, both the SUS and the EMO were significantly predictive of overall experience and likelihood-to-recommend. When modeled together, however, the combination was only marginally more predictive for overall experience and not at all for likelihood-to-recommend compared to SUS-only modeling. A subsequent analysis by Lewis et al. (2015) of data from an unmoderated usability study found that, along with key EMO factors, the SUS continued to contribute significantly to the outcome metrics of satisfaction and likelihood-to-recommend.

4.4. Relationship between SUS and other user experience metrics

The SUS is a metric of perceived usability, which is a key aspect of user experience. As in the research described in the previous section by Lewis, Brown, & Mayes. (2015), it is important to include other metrics in a battery designed to assess the broader construct of user experience. Future research should continue investigating the relationships between the SUS (and other measures of perceived usability) with metrics such as visual appeal, usefulness, and trust on outcome metrics like overall experience and loyalty metrics like likelihood-to-recommend (e.g., see the items included in the SUPR-Q, Sauro, 2015).

4.5. Minimized version of the SUS

Lewis and Sauro (2017a) demonstrated that removing any individual item from the SUS had no appreciable effect on the resulting score (given appropriate adjustment of the computational formula). Lah and Lewis (2016) reported a similar outcome when removing a specific pair of items (Items 2 and 8). No one has yet systematically investigated the effect of removing all possible pairs of items from the SUS. Carrying this line of research even further, it would be interesting to see if there is a two- or three-item version of the SUS that closely corresponds to the standard SUS. Such a minimized version of the SUS might prove to be a better surrogate measurement than the UMUX-related metrics, although, given how well they seem to correspond with concurrently collected SUS means, it might be difficult to compete with them.

5. Conclusions

Despite its humble beginnings, the SUS has become the most widely used measure of perceived usability, and is likely to remain so for the foreseeable future. Research into its psychometric properties (reliability, validity, and sensitivity) has been universally favorable. Over the past 10 years, there has been an explosion of research, starting with Bangor et al. (2008), which has led to several ways to assess the magnitude of SUS means through comparison with norms (e.g., Sauro–Lewis CGS, Sauro & Lewis, 2016; ratings of everyday products, Kortum & Bangor, 2013), recently extended to a method for setting benchmarks for individual SUS items (Lewis & Sauro, In press).

The controversy that began in 2009 (Lewis & Sauro) regarding the factor structure of the SUS has likely been resolved by a recent very-large-sample study (Lewis & Sauro, 2017b). Originally designed to be a unidimensional measure of perceived usability, the SUS appears to have a bidimensional structure that matches item tone, with the positive-tone items aligning with one factor and negative-tone items with the other. Because this artifactual structure is of little interest when measuring perceived usability, the pragmatic practice is to treat it as originally intended, as a unidimensional measure with no other underlying structure of interest.

The SUS has also proven to be a very flexible questionnaire, unaffected by minor wording changes (Bangor et al., 2008). Research indicates that even as extreme a manipulation as rewording its negative-tone items to create a positive-tone version does not dramatically affect the resulting scores (Sauro & Lewis, 2011). For situations in which practitioners need to drop an item, they can do so with confidence that the resulting scores from all 9-item version will not differ appreciably from the standard 10-item version (Lewis & Sauro, 2017a).

There have been a number of published translations of the standard SUS, with some psychometric findings presented for Arabic (AlGhannam, Albustan, Al-Hassan, & Albustan, 2017), Slovene (Blažica & Lewis, 2015), Polish (Borkowska & Jach, 2016), Italian (Borsci et al., 2009), Persian (Dianat, Ghanbari, & AsghariJafarabadi, 2014), and Portuguese (Martinsa et al., 2015). International usability and user experience research would benefit from the publication of additional translations. The translations appear to have similar psychometric properties as the English version, but there are numerous gaps that, hopefully, future research will fill.

Beyond mere correlation, recent research shows that the magnitude of SUS means closely correspond with means of other questionnaires designed to assess perceived usability. This has been demonstrated for the CSUQ (Lewis, 2018) and UMUX-related metrics (Finstad, 2010, 2015; Lewis et al., 2013). A practical consequence of this is that researchers or practitioners who work with one or more of these alternate measures can, with appropriate caution, use published SUS norms to interpret the scores of the alternate measures. Future research in correspondence should extend this work to similar metrics from other fields (e.g., TAM) and to metrics that attempt to go beyond perceived usability to a broader measure of user experience.

The question of which version of the SUS to use depends on the researcher's needs. When there is a process in place to control the types of response and scoring errors associated with the standard version (as shown in Figure 1, with "awkward" in place of Brooke's original "cumbersome"), it is reasonable to use the standard version for consistency with the majority of the published research, especially given that this is the research on which existing SUS norms have been based. When these processes are not in place or would be impossible to implement (e.g., unmoderated remote usability studies or surveys), researchers should consider using the positive version (as shown in Figure 2). If there is a desire to use the ultra-short UMUX-LITE in place of the SUS, practitioners should collect enough concurrent data to verify that the SUS and UMUX-LITE (with or without regression adjustment) means correspond closely in their contexts of measurement.

Using an evolutionary analogy, the SUS appears to have risen to the top of the food chain for metrics of perceived usability. Unless there is a compelling reason to use one of its competitors (e.g., 2-item UMUX-LITE when the 10-item SUS would require too much space on a survey; the SUPR-Q for professionally curated norms), researchers and practitioners should strongly consider using the SUS when measuring perceived usability.

ORCID

James R. Lewis http://orcid.org/0000-0002-3295-5392

References

AlGhannam, B. A., Albustan, S. A., Al-Hassan, A. A., & Albustan, L. A. (2017). Towards a standard Arabic System Usability Scale (A-SUS): Psychometric evaluation using communication disorder app. *International Journal of Human-Computer Interaction*. doi:10.1080/10447318.2017.1388099

Anastasi, A. (1976). *Psychological testing*. New York, NY: Macmillan.

Bangor, A., Joseph, K., Sweeney-Dillon, M., Stettler, G., & Pratt, J. (2013). Using the SUS to help demonstrate usability's value to business goals. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 202–205). Santa Monica, CA: HFES.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*, 574–594.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Barnette, J. J. (2000). Effects of stem and Likert response reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370.

Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. Paper presented at the Usability Professionals Association Annual Conference, UPA, Orlando, FL.

Berkman, M. I., & Karahoca, D. (2016). Re-assessing the Usability Metric for User Experience (UMUX) Scale. *Journal of Usability Studies*, *11*(3), 89–109.

Blažica, B., & Lewis, J. R. (2015). A Slovene translation of the System Usability Scale (SUS). *International Journal of Human-Computer Interaction*, *31*, 112–117.

Borkowska, A., & Jach, K. (2016). Pre-testing of polish translation of System Usability Scale (SUS). In J. Świątek, Z. Wilimowska, L. Borzemski, & A. Grzech (Eds.), *Proceedings of 37th International Conference on Information Systems Architecture and Technology – ISAT 2016 – Part I* (pp. 143–153). New York, NY: Springer.

Borsci, S., Federici, S., Gnaldi, M., Bacci, S., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: An exploratory analysis of SUS, UMUX and UMUX-LITE. *International Journal of Human-Computer Interaction.*, *31*, 484–495.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive Processes*, *10*, 193–197.

Borsci, S., Federici, S., Mele, M. L., & Conti, M. (2015). Short scales of satisfaction assessment: A proxy to involve disabled users. In M. Kurosu (Ed.), *Proceedings of HCII 2015* (pp. 35–42). Los Angeles, CA: Springer.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London, UK: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In E. Soloway, D. Frye, & S. B. Sheppard (Eds.), *Proceedings of CHI 1988* (pp. 213–218). Washington, DC: Association for Computing Machinery.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, California: Harcourt Brace Jovanovich.

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: A review and program. *Educational and Psychological Measurement*, *48*, 687–693.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Experimental Psychology*, *78*(1), 98–104.

Davis, D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319–339.

Dianat, I., Ghanbari, Z., & AshgariJafarabadi, M. (2014). Psychometric properties of the Persian language version of the System Usability Scale. *Health Promotion Perspectives*, *4*(1), 82–89.

Diefenbach, S., Kolb, N., & Hassenzahl, M. (2014). The "Hedonic" in human-computer interaction: history, contributions, and future research directions. In *Proceedings of the 2014 Conference on Designing Interactive Systems – DIS 14* (pp. 305–314). New York, NY: Association for Computing Machinery.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Finstad, K. (2006). The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*, 323–327.

Göransson, B. (2011). *SUS Svensk: System Usability Scale in Swedish*. Retrieved from http://rosenfeldmedia.com/surveys-that-work/sus-svensk-system-usability-sc/

Grier, R. A. (2013). The potential utility of the System Usability Scale in U.S. military acquisition. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 206–209). Santa Monica, CA: Human Factors and Ergonomics Society.

Grier, R. A., Bangor, A., Kortum, P. T., & Peres, S. C. (2013). The System Usability Scale: Beyond standard usability testing. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 187–191). Santa Monica, CA: Human Factors and Ergonomics Society.

ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11, Guidance on usability (ISO 9241-11:1998E)*. Geneva, Switzerland: Author.

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of websites. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 424–428). Santa Monica, CA: HFES.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*, 210–212.

Kortum, P., & Acemyan, C. Z. (2013). How low can you go? Is the System Usability Scale range restricted? *Journal of Usability Studies*, *9*(1), 14–24.

Kortum, P., & Acemyan, C. Z. (2018). The impact of geographic location on the subjective assessment of system usability. *International Journal of Human-Computer Interaction*. doi:10.1080/10447318.2018.1437865

Kortum, P., & Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale. *International Journal of Human-Computer Interaction*, *29*, 67–76.

Kortum, P., & Johnson, M. (2013). The relationship between levels of user experience with a product and perceived system usability. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 197–201). Santa Monica, CA: Human Factors and Ergonomics Society.

Kortum, P., & Oswald, F. L. (2017). The impact of personality on the subjective assessment of usability. *International Journal of Human-Computer Interaction*, *34*(2), 177–186.

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction*, *31*(8), 518–529.

Lah, U., & Lewis, J. R. (2016). How expertise affects a digital-rights-management sharing application's usability. *IEEE Software*, *33*(3), 76–82.

LaLomia, M. J., & Sidowski, J. B. (1990). Measurements of computer satisfaction, literacy, and aptitudes: A review. *International Journal of Human–Computer Interaction*, *2*, 231–253.

Landauer, T. K. (1997). Behavioral research methods in human-computer interaction. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 203–227). Amsterdam, Netherlands: Elsevier.

Lewis, J. R. (1990). *Psychometric evaluation of a Post-Study System Usability Questionnaire: The PSSUQ (Tech. Report 54.535)*. Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (1992). Psychometric evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1259–1263). Santa Monica, CA: Human Factors Society.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*, 463–488.
Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*. doi:10.1080/10447318.2017.1418805
Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated usability study. *International Journal of Human-Computer Interaction*, *31*(8), 545–553.
Lewis, J. R., & Mayes, D. K. (2014). Development and psychometric evaluation of the Emotional Metric Outcomes (EMO) questionnaire. *International Journal of Human-Computer Interaction*, *30*(9), 685–702.
Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. *In* M. Kurosu (Ed.), *Human centered design* (pp. 94–103). Heidelberg, Germany: Springer-Verlag.
Lewis, J. R., & Sauro, J. (2017a). Can I leave this one out? The effect of dropping an item from the SUS. *Journal of Usability Studies*, *13*(1), 38–46.
Lewis, J. R., & Sauro, J. (2017b). Revisiting the factor structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183–192.
Lewis, J. R., & Sauro, J. (In press). Item benchmarks for the System Usability Scale. *To Appear in Journal of Usability Studies, 13*(3).
Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE—When there's no time for the SUS. In *Proceedings of CHI 2013* (pp. 2099–2102). Paris, France: Association for Computing Machinery.
Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505.
Lindgaard, G., & Kirakowski, J. (2013). Introduction to the special issue: The tricky landscape of developing rating scales in HCI. *Interacting with Computers*, *25*, 271–277.
Lucey, N. M. (1991). *More than meets the I: User-satisfaction of computer systems* (Unpublished thesis for Diploma in Applied Psychology). University College Cork, Cork, Ireland.
Martinsa, A. I., Rosa, A. F., Queirós, A., & Silva, A. (2015). European Portuguese validation of the System Usability Scale. *Procedia Computer Science*, *67*, 293–300.
McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on System Usability Scale ratings. *Journal of Usability Studies*, *7*(2), 56–67.
McSweeney, R. (1992). *SUMI: A psychometric approach to software evaluation* (Unpublished M.A. (Qual.) thesis in applied psychology). University College of Cork, Cork, Ireland.
Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*, 396–402.
Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the System Usability Scale (SUS): SUS in the wild. In *Proceedings of the Human Factors and Ergonomics Society 57th Meeting* (pp. 192–196). Santa Monica, CA: HFES.
Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a Computer Anxiety Scale. *Educational and Psychological Measurement*, *50*, 603–610.
Reichheld, F. (2003). The one number you need to grow. *Harvard Business Review*, *81*, 46–54.
Reichheld, F. (2006). *The ultimate question: Driving good profits and true growth*. Boston, MA: Harvard Business School Press.
Rummel, B. (2015). *System Usability Scale – Jetzt auch auf Deutsch*. Retrieved November 27, 2017, from https://experience.sap.com/skillup/system-usability-scale-jetzt-auch-auf-deutsch/
Sauro, J. (2010). *That's the worst website ever! Effects of extreme survey items*. Retrieved March 24, 2011, from www.measuringu.com/blog/extreme-items.php
Sauro, J. (2011). *A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC.
Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86.
Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of CHI 2009* (pp. 1609–1618). Boston, MA: Association for Computing Machinery.
Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of CHI 2011* (pp. 2215–2223). Vancouver, Canada: ACM.
Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research* (1st ed.). Waltham, MA: Morgan Kaufmann.
Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan Kaufmann.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.
Schmitt, N., & Stuits, D. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373.
Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*(4), 1101–1114.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.
Stewart, T. J., & Frye, A. W. (2004). Investigating the use of negatively-phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, *79*(10 Supplement)), S1–S3.
Tossell, C. C., Kortum, P., Shepard, C., Rahmati, A., & Zhong, L. (2012). An empirical analysis of smartphone personalization: Measurement and user variability. *Behaviour & Information Technology*, *31*(10), 995–1010.
Tullis, T. S., & Albert, B. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Burlington, MA: Morgan Kaufmann.
Tullis, T. S., & Stetson, J. N. (2004). *A comparison of questionnaires for assessing website usability*. Paper presented at the Usability Professionals Association Annual Conference. UPA, Minneapolis, MN. Retrieved September 13, 2017, from https://www.researchgate.net/publication/228609327_A_Comparison_of_Questionnaires_for_Assessing_Website_Usability
Wong, N., Rindfleisch, A., & Burroughs, J. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, *30*, 72–91.

About the Author

James R. (Jim) Lewis is a user experience practitioner at IBM who has published influential papers on usability testing and measurement. He is an IBM Master Inventor with 91 US patents issued to date. His books include*Practical Speech User Interface Design and (with Jeff Sauro) Quantifying the User Experience*.